

APPLICATION
FOR
UNITED STATES LETTERS PATENT

TITLE: DETERMINING PACKET SIZE IN NETWORKING
APPLICANT: JAMES L. JASON JR.

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EL688321088US

I hereby certify that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

October 22, 2001

Date of Deposit


Signature

Mike Augustine

Typed or Printed Name of Person Signing Certificate

DETERMINING PACKET SIZE IN NETWORKING

TECHNICAL FIELD

This invention relates to determining packet size in networking.

BACKGROUND

When communicating a message between two points on a computer network, the message can be sent in discrete-sized packets. Properties of the network constrain the maximum size of the packet, or maximum transfer unit (MTU), which can be sent along a network path 20 from a particular sending point 22 to a particular receiving point 24, as shown in Figure 1. For example, the underlying hardware used to implement the path on the network, such as hardware using 100 Mbit Ethernet technology, will impose limitations on the path MTU. Furthermore, the data packets may pass through several intermediary points 26, such as routers, as they travel from the sending point 22 to the receiving point 24. The path MTU may be further limited by the technology used between some of these intermediary points.

Figure 2a shows a typical Internet Protocol (IP) data packet. Prepended to the data 301 is an IP header 303, containing information necessary for communicating the packet from the sending point 22 to the receiving point 24. Figure 2b shows an IP datagram that has been encapsulated with an additional outer IP header 34. This additional encapsulation can take place at

the sending point 22 or one of the routers. Encapsulating, with an outer header, data that has previously been encapsulated with an inner header is commonly referred to as IP-in-IP encapsulation, or tunneling.

5 If the sending point 22 sends a packet that is larger than the path MTU, routers along the network path will fragment the packet 28 into smaller pieces, or fragments 29, during the transmission. Typical fragments 29 are shown in Figure 2c. After creating the fragments, the router re-encapsulates the data 10 such that each of the fragments 29 will have the tunneling outer header 34 prepended to the data 30, but only the first fragment 29a of the data packet will have the inner header 32. These fragments 29 are cached at the receiving point 24 until all of the fragments 29 of the packet have been received or until the 15 reassembly timer for the datagram has expired and the fragments are discarded. The information in the headers 32, 34 gives the receiving point information on how the fragments 29 should be grouped and handled upon receipt. As shown in Figure 2d, the fragments 29 can then be reassembled (into the packet 28) from the cache at the receiving point 24. After reassembly, the 20 tunneling outer IP header can be stripped off, leaving the inner packet 31 with only the inner IP header prepended. The packet 31 can then be sent on in the usual way.

One way of eliminating the consumption of computing 25 resources needed for caching and reassembly, for example, in the tunneling context, is for the sending point 22 to determine the path MTU in advance. With knowledge of the path MTU, the sending

point 22 can send packets 52, shown in Figure 2e, which will be small enough so that they will not be fragmented in their travel to the destination. Because a router does not fragment the packets 52, they will not need to be cached and reassembled at 5 the receiving point 24. The router simply has to remove the outer IP header and send the encapsulated datagram on its way. With reference to Figure 3, the sending point 22 begins the determination by sending a probe packet in which a ``don't fragment'' bit is set (steps 300 - 304). The size of the probe 10 packet is the largest possible packet that the networking technology at the sending point will allow (the MTU of the link layer). Because the ``don't fragment'' bit is set, if the packet is larger than the MTU of the path, it will not be fragmented. Instead, an error message will be sent back to the sending point 15 if that packet otherwise would have been fragmented (steps 306 - 308). The sending point then sends a smaller probe packet with the ``don't fragment'' bit set. This process (steps 300 - 308) is repeated until a packet is sent that is small enough to travel to the receiving point without fragmenting. When it receives no 20 error message, the sending point knows that the size of the message that was able to pass is the path MTU (step 310).

DESCRIPTION OF DRAWINGS

Figure 1 is a block diagram of a network path;
Figures 2a - 2e are diagrams of data packets;
25 Figure 3 is a flow chart of a method of determining a path MTU at a sending point;

Figure 4 is a diagram of a computer network;

Figure 5 is a flow chart of a method of determining a path MTU at a receiving point;

Figure 6 is a flow chart of a more detailed method of
5 determining a path MTU at a receiving point; and

Figure 7 is a block diagram of an interface device.

DETAILED DESCRIPTION

To eliminate the need for sending several probe packets to determine a path MTU, the MTU is determined at a receiving network point and communicated to the sender. Figure 4 shows a computer network 36 for communicating packets of data. The network 36 includes a sending interface device 38 and a receiving interface device 40. The two interface devices 38, 40 are connected to one another across a sub network 42. A sending computer 44 on sub network 46 communicates with a receiving computer 48 on sub network 50 by sending data packets through the sending interface device 38. The sending interface device 38 regulates the size of the data packets and encapsulates them with header information. Each packet is then sent to the receiving interface device 40, which decapsulates the packet before it is sent to the receiving computer 48 on the sub network 50.

Figure 5 shows a method of determining the path MTU at the receiving point. When the sending computer sends a packet to the receiving computer, the sending interface device assumes that the path MTU is the MTU of the link layer at the sending interface device and the sub network. It then sends a packet of that size

(step 500). If that packet is larger than the actual MTU of the path between the sending and receiving interface devices, the packet gets fragmented (step 502). The receiving interface device receives the fragments and determines the size of the largest 5 fragment (step 504). This size is the path MTU (step 506). The receiving interface device communicates this MTU to the sending interface device (step 508). The sending interface device can then optimize the efficiency of data communication between the two interface devices by sending packets of the largest possible 10 size that will not get fragmented (step 510).

Figure 6 shows a more detailed method of determining the path MTU at the receiving point. The interface devices provide interfaces between multiple sub networks, encapsulate and decapsulate data, and collect information about the data passing 15 through them. The sending interface device 38 initially assumes that the MTU for the border between the sending interface device and the sub network 42 is the MTU of its link layer (step 600). The sending interface device 38 reports this information to a sending policy broker 52 (step 602). Similarly, other interface 20 devices on the network, including the receiving interface device 40, assume an initial MTU for their interfaces and report this information to their corresponding policy brokers. The brokers then exchange this information among themselves (step 606). Each broker then distributes this information to its corresponding 25 interface device (608). Based on this information, the sending interface device 38 assumes a path MTU between it and the receiving interface device (step 610). It stores this assumed

MTU in a computer memory 54. Similarly, the receiving interface device 40 assumes a path MTU and stores it in a computer memory 56. Alternatively, each interface device can assume that the path MTU is the MTU of the link layer at the border between that interface device and the sub network without incorporating information collected by brokers. The actual path MTU may be different than either of the assumed MTUs due to network constraints not factored into the initial exchange of information between the brokers and interface devices.

The sending computer 44 sends data to the receiving computer 48 through the sending interface device 38 (step 612). The sending interface device 38 breaks up the data and encapsulates it to form packets of the size of the assumed MTU stored in the computer memory 54. If the packet is larger than the path MTU, the packet is fragmented as it is sent to the receiving interface device 40 (step 614). After receiving the packets, the receiving interface device 40 analyzes the fragments to determine their sizes (616). If the fragment being analyzed is the last fragment in a packet (step 618), the size is checked to see if it is greater than the path MTU (as are non-fragmented datagrams). If so, the path MTU is changed. If it is not larger than the path MTU, then the path MTU is not changed as it most likely that the last fragment will be smaller than the path MTU.

If the fragment is not the final fragment, then its size is compared to the assumed path MTU stored in the computer memory 56 (step 622). If it is the same size as the receiving interface device's assumed path MTU, then the receiving interface device 40

will consider the assumed path MTU to be the actual path MTU and will not change its assumed path MTU (step 620). If the fragment is larger than the assumed path MTU, the receiving interface device 40 will know that the actual path MTU is greater than the assumed path MTU and change the assumed path MTU stored in the memory 56 to be equal to the fragment size (step 624). If the fragment is smaller than the assumed path MTU, the receiving interface device 40 will know that the actual path MTU is smaller than the assumed path MTU and change the assumed path MTU in the memory to be equal to the fragment size (step 624).

If the size of the packet is not larger than the path MTU, it will not be fragmented when it is communicated to the receiving interface device 40. The receiving interface device 40 analyzes the size of the unfragmented packet and compares it to the assumed path MTU in the memory 56 (step 626). If it is greater than the assumed path MTU, the path MTU is changed in the memory 56 to equal the size of the packet (step 624), since packets of at least that size can be sent by the sending interface device 38 without fragmentation. If it is not greater than the assumed MTU, the assumed path MTU is not changed in the memory 56 (step 620).

After analyzing the packet or fragments, the receiving interface device 40 reports its assumed path MTU to a receiving broker 58 (step 628). Alternately, the interface device 40 only reports the assumed path MTU to its broker 58 if its assumed MTU has changed. In either case, the receiving broker 58 communicates the assumed path MTU to the sending broker (step

630). The sending broker 52 communicates the assumed path MTU to the sending interface device 38 (step 632), which updates its assumed path MTU in the memory 54. In subsequent communications to the receiving interface device 40, the sending interface device 38 sends packets of the size of the new assumed path MTU.

5 The network path between the sending interface device and the receiving interface device may not remain static. It is possible that segments of the network path connecting intermediary points between the sending and receiving interface devices could be broken, or shorter or more efficient segments could be added. This changes the topology of the network and could change the path that data packets travel when being communicated between the sending and receiving interface devices. Thus the path MTU between the sending and receiving interface devices may occasionally change. One way to compensate for this change is for the receiving interface device 40 to communicate a new path MTU to the sending interface device 38 any time it detects a change. Another way is for the sending interface device 38 to occasionally send a control packet to the receiving interface device 40. This packet is the largest possible packet allowed by the technology of the sending interface device's link layer. As above, if the packet is larger than the actual path MTU, it will be fragmented before reaching the receiving interface device 40. The receiving interface device 40 then analyzes the packet or fragments to determine the actual path MTU, updates the assumed path MTU in the memory 56, and reports it back to the sending interface device 38, which updates the

assumed MTU value in the memory 54. The sending interface device 38 will send subsequent communications in packets of the size of the new assumed path MTU until that value is again changed in the memory 54.

5 Figure 7 shows an interface device. A data message 62 enters the interface device 60 and is classified using a data classification module 64. The data classification module 64 analyzes the header encapsulated with the data to determine whether the data is a packet or a fragment, and if it is a fragment, to determine whether it is the last fragment of a packet.
10 The data can be classified using a variety of criteria to determine how the network prioritizes and processes the data. A policy, including information about the path MTU, is dictated to the interface device 60 by a broker 68 corresponding to the
15 interface device 60, and is received through a remote policy interface 70. The classification module analyzes the data and determines whether it needs to be encapsulated or decapsulated. The encapsulation or decapsulation is then performed, according to the policy, using a packet manipulation module 72.

20 In the case of encapsulation, the MTU value, which is known to the packet manipulation module, is used to fragment the inner packet as shown in Figure 2e with each fragment 52 carrying the related inner IP header 32. The tunneling outer header 34 is then prepended to each fragment. The data packets are then
25 queued and scheduled for sending according to a policy, using a queuing and scheduling module 74. The policy is received from the broker through the remote policy interface 70.

Other embodiments are within the scope of the following claims.

10559-502001 / P11794